Survival analysis of pension scheme mortality when data are missing

Heriot-Watt University - Actuarial Mathematics and Statistics

Survival Analysis of Pension Scheme Mortality when Data are Missing

Francesco Ungolo, Marcus C. Christiansen,
Torsten Kleinow, Angus S. Macdonald

January 16, 2019

Missing data is a problem that may be faced by actuaries when analysing mortality data. In this paper we deal with pension scheme data, where the future lifetime of each member is modelled by means of parametric survival models incorporating covariates, which may be missing for some individuals. Parameters are estimated by likelihood-based techniques. We analyse statistical issues, such as parameter identifiability, and propose an algorithm to handle the estimation task.

Finally, we analyse the financial impact of including covariates maximally, compared with excluding parts of the mortality experience where data are missing; in particular we consider annuity factors and mis-estimation risk capital requirements.

Keywords: Mortality, Survival model, Longevity risk, Missing data, Mortality models with covariates

1. Introduction

Our motivating example is that of an actuary modelling the mortality of a pension scheme. We assume data is available at the level of the individual scheme member, including date of birth, date of joining the scheme, date of leaving observation, and reason for cessation of observation (usually, death or right censoring).

Often, other information will be available in the form of covariates. Gender is usually known, but increasingly often much richer information is available, such as benefit or pension size, geo-demographic profile, or health status. Knowing how mortality depends on these can be important, especially if risk may be concentrated on certain subsets of the membership (for example, persons with the largest pensions may have the lightest mortality Richards [2008]).

The modelled mortality may be used for certain purposes that are statistical in nature, such as estimating risk reserves. Then the estimation error incurred in fitting the model may be material. If we fit a parametric model of the hazard rate, for example, the parameters estimates have sampling error, which is inherited by any quantity calculated using the estimated hazard rates.

In principle, estimation error can be reduced by modelling a larger population, if it remains sufficiently homogeneous. So, in the UK, the Continuous Mortality Investigation (CMI) often combines data provided by different life insurers writing the same class of business. Our pensions actuary may wish, similarly, to combine the data from two or more pension schemes, to obtain a larger and, it is hoped, more reliable dataset. But what then happens, if statistically significant covariates are observed only in some of the various pension schemes? For example, suppose Scheme A may provide benefit amount but not geo-demographic profile, while Scheme B may does the opposite.

One option, always, is to drop the covariate information and model only what the schemes have in common. But it would be useful, if possible, to model the schemes jointly while retaining what covariate information there is, since the former option could imply problems of model misspecification which may cause lack of consistency of the maximum likelihood estimator, see White [1982]. That is the problem considered here.

An approach we will find useful is to regard the combined pension schemes as a single dataset in which the absent covariate information is treated as missing data. There is a large statistical literature on missing data, a problem dated back to the works of Wilks [1932] and Lord [1955] which focus on the normal distribution, while in the field of survival data Schluchter and Jackson [1989] consider a log-linear model for the analysis of censored survival data with categorical covariates, and the work of Herring and Ibrahim [2001] focuses on the Cox proportional hazards model. In these two studies a general missing data pattern is considered and parameters are fitted using the EM algorithm.

The remainder of this paper develops as follows: Section 2 defines the basic survival model, Section 3 defines the notation for the probability distribution of the covariates, and Section 4 describes the likelihood inferential framework and its extension to the case when data are missing.

We also carry out an empirical study where we consider a dataset for the male members aged 60 and above of a UK pension scheme, which includes survival times and covariates for each individual. Section 5 describe this dataset, a preliminary data analysis and our chosen parametric form for the force of mortality.

Using our pension fund data with complete observations we first simulate datasets with missing data by removing covariates for some units (Section 6). Hence, we describe the estimation problem, the mathematical conditions for identifiability of parameter estimates and the fitting algorithm.

When considering the case of two statistically significant covariates which are never jointly observed, the model is very likely to be not identifiable given the data. For this reason, in Section 7 we consider the case where auxiliary information, in the form of a very small sample of lives whose covariates are completely observed, is available in order to overcome the identifiability issue. In actuarial applications, obtaining a small set of data where both covariates are jointly observed and no data is missing might be very expensive. This may be compared to the de-duplication problem faced by the Continuous Mortality Investigation (CMI) of the IFoA. For the CMI, data on policies rather than lives are collected which leads to a problem of duplicates — one person having several policies. To assess the impact of duplicates on mortality projections, special investigations to estimate the distribution of the number of duplicates policies were carried out. Those were also based on ancillary information.

In Section 8 we discuss the financial impact of combining different datasets for estimating the hazard rates, and Section 9 concludes.

2. Survival Model

The future lifetime of an individual currently aged x is a random variable and we denote by f_x, F_x and $S_x = 1 - F_x$ the density function, the distribution function and the corresponding survival function of this random variable. As usual we denote by μ_x the force of mortality at age x, which is given by

$$\mu_x = \frac{f_0(x)}{1 - F_0(x)}. \tag{2.1}$$

We then obtain the following two well-known results, see for example, Macdonald [1996]:

$$S_x(t) = \exp\left\{ -\int_0^t \mu_{x+s} ds \right\} \tag{2.2}$$

and

$$f_x(t) = S_x(t)\mu_{x+t} = \exp\left\{ -\int_0^t \mu_{x+s} ds \right\} \mu_{x+t}. \tag{2.3}$$

Our statistical model for the future lifetime of an individual will be based on a model for the force of mortality μ_x. We assume that $\mu_x = \mu_x(z;\tau)$ where z is a vector of explanatory variables (covariates) which do not depend on age, and τ is an unknown parameter vector. We then also use the notation $S_x(t \mid z;\tau)$ and $f_x(t \mid z;\tau)$ for the functions in (2.2) and (2.3).

3. Covariates as Random Variables

While in regression analysis covariates are often considered to be deterministic, we find it helpful to model them as random variables.

We denote by Z the p-dimensional random vector indicating the set of covariates of a randomly chosen individual whose realized value z determines the individual's force of mortality $\mu_x(z;\tau)$. The distribution function of Z is denoted by $F_Z(z;\zeta)$ where ζ is an unknown parameter vector, and $f_Z(z;\zeta)$ is the corresponding density or probability function.

4. Observations and Likelihood function

4.1. Complete Observations

Our aim in this section is to derive the joint likelihood function for estimating the parameter vector (τ, ζ).

We assume that n individuals are observed during a finite time period. We denote by

- x_i the age of individual i at the start of the observation period;
- t_i the realized value of a random variable T_i describing the total time that individual i is observed and alive during the observation period (T_i is often called the observed exposure);
- d_i an indicator, the realized value of a random variable D_i, equal to 1 if individual i is observed to die, and 0 if observation ends by censoring; and
- z_i the realized value of a random variable Z_i describing the covariates for this individual.

During a finite observation period some individuals will die while others will survive (right censoring). For those who die, T_i is the remaining lifetime from age x_i, while for those who survive, all that we observe is that their remaining lifetime is greater than T_i.

We assume that $Z_1, \ldots, Z_n$ are independent and identically distributed (i.i.d.), and that conditional on their realized values and the observed ages $x_1, \ldots, x_n$ the observed exposures $T_1, \ldots, T_n$ are also i.i.d. Also, we assume that the censoring is non-informative, see Macdonald et al. [2018].

To begin with we suppose that observation is complete, in the sense that every covariate vector z_i is fully observed.

We denote by L_i the contribution to the likelihood function from individual i. The total likelihood is then $L = \prod L_i$.

For those individuals observed to die ($d_i = 1$) the contribution to the likelihood is given by the conditional density of T_i given x_i and z_i multiplied by the density/probability function of z_i, that is, $L_i(\tau, \zeta) = f_{x_i}(t_i \mid z_i; \tau) f_Z(z_i, \zeta)$. In what follows we define the joint density of (T, Z) as the full model (given the age X), while the density of T conditional on Z is called the partial model.

For those individuals whose observed lifetimes are right-censored ($d_i = 0$) the exact time of death is not observed and the contribution to the likelihood is $L_i(\tau, \zeta) = S_{x_i}(t_i, z_i; \tau) f_Z(z_i; \zeta)$.

We combine those two cases in the usual way and obtain

$$L_i(\tau,\zeta) = \exp\left\{-\int_0^{t_i} \mu_{x_i+s}(Z_i;\tau)ds\right\} \mu_{x_i+t_i}(z_i;\tau)^{d_i} f_Z(Z_i;\zeta). \tag{4.1}$$

When all elements of Z are observed for all individuals, this likelihood factorizes into a function of τ and a function of ζ:

$$\begin{aligned} L(\tau,\zeta) &= \prod_i L_i(\tau,\zeta) \\ &= \underbrace{\prod_i \exp\left\{-\int_0^{t_i} \mu_{x_i+s}(z_i;\tau)ds\right\} \mu_{x_i+t_i}^{d_i}(z_i;\tau)}_{L^\tau(\tau)} \underbrace{\prod_i f_Z(z_i;\zeta)}_{L^\zeta(\zeta)}. \end{aligned} \tag{4.2}$$

In this case τ can be estimated independently of ζ, and the distribution of the covariates is usually disregarded in regression analysis. However, we will see in the next section that this is not possible when some observations of the Z_i are missing.

4.2. Missing Observations

In this paper we focus on the situation in which not all components of the covariates Z are observed for some individuals.

Considering the p-dimensional vector $Z_i = (Z_{i,1}, \ldots, Z_{i,p})$ we form the two vectors Z_i^{obs} and Z_i^{mis} where Z_i^{obs} contains the q_i observed components of Z_i, and Z_i^{mis} contains the remaining $p - q_i$ unobserved components (missing observations). Each observed individual can thus have a different set of observed covariates.

For each individual the density of the covariate vector $Z_i = (Z_i^{\text{obs}}, Z_i^{\text{mis}})$, now has the factorization $f_{Z_i}(z_i) = f_{Z_i^{\text{obs}}}(z_i^{\text{obs}};\zeta)\, f_{Z_i^{\text{mis}}|Z_i^{\text{obs}}}(z_i^{\text{mis}} \mid z_i^{\text{obs}};\zeta)$ (see Appendix A for a more detailed treatment).

We assume that missing covariates are missing at random, see Rubin [1976]. This means that the failing to observe a covariate does not provide any information about the value of the covariate itself.

Similarly to equation (4.1), the likelihood contribution of each individual is based on the observed lifetime T, the age X, and the set of observed covariates Z^{obs}:

$$L_i(\tau,\zeta) = \int_{\mathcal{S}_{Z_i^{\text{mis}}}} \exp\left\{-\int_0^{t_i} \mu_{x_i+s}(z_i^{obs}, y;\tau)ds\right\} \mu_{x_i+T_i}^{d_i}(z_i^{obs}, y;\tau) f_Z(z_i^{obs}, y;\zeta)dy \tag{4.3}$$

where we integrate over the state space $\mathcal{S}_{Z_i^{\text{mis}}}$ of Z_i^{mis}, and the integral over $\mathcal{S}_{Z_i^{\text{mis}}}$ becomes a sum for those components of Z_i^{mis} having a discrete state space.

In this case we cannot factorize the total likelihood into separate functions of τ and ζ as in (4.2). This means that we cannot estimate τ independently of ζ. We are forced to

estimate τ and ζ jointly based on the observed lifetimes and covariates, even though we are only interested in τ.

While we will discuss the identifiability of the parameters in a specific model in Section 6.2, we mention here that when data are missing and τ and ζ are both unknown, then the identifiability of the full model given by the joint density $f_{T,Z}(t, z \mid x; \tau, \zeta) = f_{x_i}(t \mid z; \tau) f_Z(z; \zeta)$ cannot be taken for granted (see Cole et al. [2010]).

In particular, if data are missing a statistical model may display parameter redundancy[1], and this affects its identifiability. Theorem 4 of Catchpole and Morgan [1997] shows that if a model is parameter redundant, then it is not locally identifiable[2].

Statistical models which are not parameter redundant are defined as full rank, although Catchpole and Morgan [1997] distinguish between essentially and conditionally full rank models. For essentially full rank models, the information matrix is full rank for all θ, while for conditionally full rank models the information matrix is full rank for some but not all values of θ.

If a model is full rank, the eigenvalues of the empirical information matrix evaluated at the MLE are all negative and far from zero.

As mentioned above, in Section 6.2 we will discuss the identifiability conditions for our specific model in more detail.

5. Data and Model

The empirical analysis in this paper is based on observations of the members of a medium-sized pension fund. In particular, we observe the values of two covariates for every member of the fund. We call this the *complete dataset*. From the complete dataset, we will artificially create datasets with missing observations by supposing that for some or all individuals, we fail to observe one or other covariate (or both covariates). First we briefly describe the complete dataset.

5.1. The Complete Dataset

The complete dataset has the following characteristics:

- 18,741 records of pensions in payment;
- 172,601.4 person-years of time exposed to risks;
- 4,956 observed deaths;
- period of observation: 10^{th} November 1992 – 31^{st} December 2009.
- The annual benefit amount is observed for each individual.

[1] A statistical model M indexed by the p-dimensional parameter vector θ is parameter redundant if it can be equivalently indexed by a smaller parameter vector ϑ of dimension d with $d < p$.

[2] Let $f_T(\,\cdot\,;\theta)$ be a parametric density indexed by the parameter vector $\theta \in \Omega_\theta$. A choice of parameters θ is locally identifiable if there exist a constant $\epsilon > 0$ such that there is no $\theta_1 \neq \theta_2$ such that $\| \theta_1 - \theta_2 \|_2 < \epsilon$ and $f_T(\,\cdot\,;\theta_1) = f_T(\,\cdot\,;\theta_2)$ (see Koller and Friedman [2009]).

- The geo-demographic profile is observed at individual level on the basis of MOSAIC profiler, widely used in the UK for pricing longevity swaps and bulk buy-outs (Richards [2008]).

The last two items together comprise the covariate vector z_i.

Although the benefit amount is observed in pounds, and could be treated as continuous, we treat it as categorical for two reasons. Firstly, Macdonald et al. [2018] showed how its direct use within the model does not improve the fit. Secondly, treating a continuous variable as categorical is more convenient in terms of parsimony of the modelling approach.

Furthermore, Madrigal et al. [2011] observed that a low pension amount can be a misleading indicator of individual affluence, as it could be observed because an employee was low-paid with long service, or highly-paid with short service.

For those reasons, we treat the benefit amount as a two-level categorical variable which we denote by B. This random variable has possible realisations *High* (high benefit) and *Low* (low benefit). Using cluster analysis (see Appendix B) we set the threshold between high and low benefits at £8,500 per annum.

Similarly, we treat the geo-demographic profile as a categorical variable and denote it by C. The level is determined by the postcode, therefore by the geographic area of a member's postal address. Using cluster analysis (see Appendix B) we determined three levels for C, denoted by 0, 1 and 2. Each represents a set of geo-demographic codes with level 0 indicating the most deprived areas of the UK and level 2 the least deprived.

As mentioned in Section 3, we treat $Z = (B, C)$ as a random variable, and assume it has a multinomial distribution Mult(1,6) with six possible outcomes, indexed by the parameter vector ζ whose elements are the probabilities of individual outcomes.

We now describe the features of the full dataset, propose a simple parametric model for the force of mortality and show how benefit level and geo-demographic profile improve the model fit.

5.2. Preliminary data analysis

We plot the crude death rate at single years of age, D_x/E_x, in Figure 5.1. The death rates are plotted separately for the two benefit levels (left plot) and the three groups based on the geo-demographic profile (right plot). Here D_x is the number of individuals who die between exact ages x and $x+1$, while E_x (exposed-to-risk) is the total time all individuals are observed and alive between these ages.

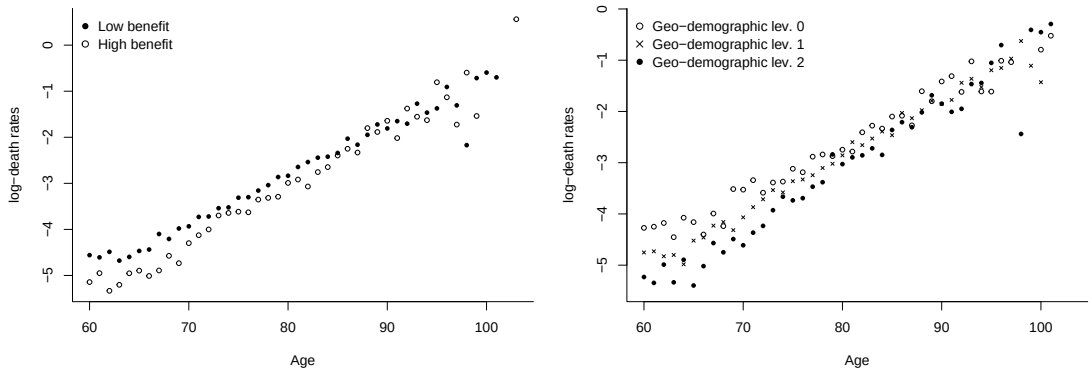

Figure 5.1: Crude death rates D_x/E_x by age for individuals with high and low benefit levels on a log scale (LHS), and crude death rates by age for individuals with different geo-demographic profiles on a log scale (RHS).

From Figure 5.1 we make three key observations.

1. There is an approximately log-linear relationship between age and death rates, even for very old ages at which sampling errors are higher because of smaller exposures. See Appendix C for full details of exposures and death counts.
2. Higher benefits are associated with lower mortality.
3. Less deprived geo-demographic status is associated with lower mortality.

The slightly irregular pattern around age 60 may be caused by our inability to distinguish between people retiring at age 60 as they had planned, and people retiring at that age because of poor health, who were willing to work for longer, as hypothesised by the CMI committee (see CMI [2007]).

5.3. Choice of Parametric Model

To keep our model simple we ignore the closing gap between mortality rates for different groups at high ages. Instead, we use a Gompertz-type model, Gompertz [1825], with group specific intercepts, that is, the force of mortality for the i-th individual is given by:

$$\mu_{x_i}(b_i, c_i; \tau) = \exp\left\{\alpha + \beta x_i + \gamma \mathbb{1}_{[b_i=\text{High}]} + \delta_1 \mathbb{1}_{[c_i=1]} + \delta_2 \mathbb{1}_{[c_i=2]}\right\} \tag{5.1}$$

where $\tau = (\alpha, \beta, \gamma, \delta_1, \delta_2)$ is the unknown parameter vector. Using (2.2) the hazard function in (5.1) corresponds to the survival function $S_{x_i}(t_i \mid b_i, c_i; \tau)$ given by:

$$S_{x_i}(t_i \mid b_i, c_i; \tau) = \exp\left\{-\left(\frac{\exp(\beta t_i) - 1}{\beta}\right) \exp\left(\alpha + \beta x_i + \gamma \mathbb{1}_{[b_i=\text{High}]} + \delta_1 \mathbb{1}_{[c_i=1]} + \delta_2 \mathbb{1}_{[c_i=2]}\right)\right\} \tag{5.2}$$

and the density function in (2.3) for the remaining life time becomes:

$$f_{x_i}\left(t_i \mid b_i, c_i; \tau\right) = S_{x_i}\left(t_i \mid b_i, c_i; \tau\right) \mu_{x_i+t_i}\left(b_i, c_i; \tau\right). \tag{5.3}$$

The parameter β measures the age effect, γ captures the effect of receiving benefits in excess of £8,500 per annum, while δ_1 and δ_2 capture the effects of different geo-demographic areas.

The parameters γ, δ_1 and δ_2 represent the mortality differentials between the different groups, and the baseline mortality $\exp(\alpha + \beta x)$, which in our model is the mortality rate of an individual aged x who has a low benefit level (up to £8,500 per annum), and who lives in the most deprived geo-demographic area.

5.4. Empirical Results from the Complete Dataset

Using the complete dataset (no missing observations) we investigate how the goodness of fit of our model for μ in (5.1) compares to the nested models with appropriate combinations of the parameters γ, δ_1 and δ_2 set to zero. To this end we consider the following models for the force of mortality $\mu_{x_i}\left(b_i, c_i; \tau\right)$:

$$\begin{array}{ll}
\text{model } M_0\text{:} & \exp\left[\alpha + \beta x_i\right] \\
\text{model } M_1\text{:} & \exp\left[\alpha + \beta x_i + \gamma \mathbb{1}_{[b_i=\text{High}]}\right] \\
\text{model } M_2\text{:} & \exp\left[\alpha + \beta x_i + \delta_1 \mathbb{1}_{[c_i=1]} + \delta_2 \mathbb{1}_{[c_i=2]}\right] \\
\text{model } M_3\text{:} & \exp\left[\alpha + \beta x_i + \gamma \mathbb{1}_{[b_i=\text{High}]} + \delta_1 \mathbb{1}_{[c_i=1]} + \delta_2 \mathbb{1}_{[c_i=2]}\right].
\end{array}$$

Model M_0 is a Gompertz model without covariates. In models M_1 and M_2 only one covariate is used (benefit amount and geo-demographic profile, respectively), and in model M_3 both covariates are used.

As mentioned earlier, with the complete dataset the likelihood function factorises as in (4.2). Since we want to estimate τ, we need to consider only the factor $L^\tau(\tau)$ and ignore $L^\zeta(\zeta)$. We compare the models using the Bayesian Information Criterion (BIC), Schwarz [1978], equal to $-2\ell^\tau(\tau) + a_J \log n$, where $\ell^\tau(\tau) = \log\left(L^\tau(\tau)\right)$. The empirical results obtained for the complete dataset are shown in Table 5.1.

Table 5.1: Parameters estimates, log-likelihood value and BIC for the four models fitted to the complete dataset.

Model	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\gamma}$	$\widehat{\delta}_{\mathbf{1}}$	$\widehat{\delta}_{\mathbf{2}}$	$\ell^\tau(\tau)$	**BIC**
M_0	−11.58	0.11	–	–	–	−20592.61	41204.90
M_1	−11.54	0.11	−0.28	–	–	−20561.88	41153.27
M_2	−11.42	0.11	–	−0.24	−0.48	−20528.93	41097.21
M_3	−11.41	0.11	−0.21	−0.22	−0.43	−20512.69	41074.57

We find that including B and C separately and jointly improves the fit of the model, with the larger part of the improvement accounted for by the geo-demographic profile;

the BIC is smallest for model M_3 in which both B and C are included. We conclude that there are significant mortality differentials between different socio-economic groups.

Information criteria compare the goodness of fit of competing models but do not tell us how well the best-fitting model explains the observed data. To investigate the quality of fit of our model in (5.1), in Figure 5.2 we plot the Poisson Deviance Residuals (Cox and Miller [1977]), for the models M_0 (LHS) and M_3 (RHS). We observe that the residuals are centred around zero and do not show any particular pattern. The two residuals which lie outside the range $[-1.96, 1.96]$ are those at ages 60 and 61 which we already discussed in Section 5.2.

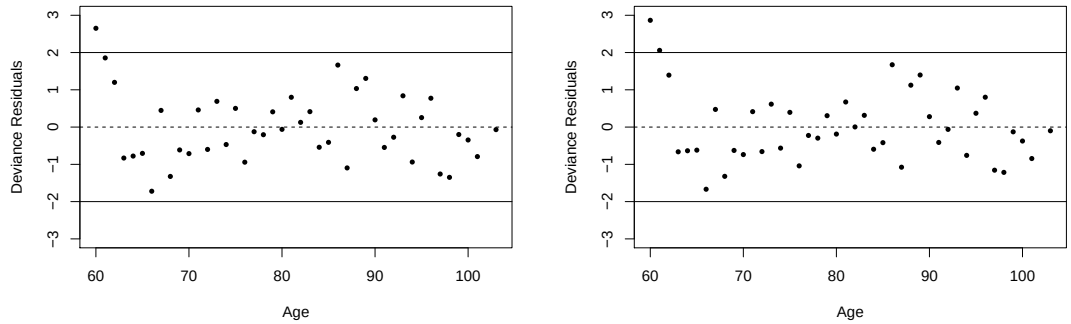

Figure 5.2: Poisson Deviance residuals plotted against the age for model M_0 (LHS), and and for model M_3 (RHS).

Given the complete dataset, the estimated joint distribution of the covariates B and C is given by the relative frequencies which are reported in Table 5.2.

Table 5.2: Empirical joint distribution of benefit level and geo-demographic profile among 18,741 pension scheme members.

		Geo-demographic profile (C)			
		0	**1**	**2**	
Benefit level (B)	**Low**	0.19	0.43	0.13	0.75
	High	0.03	0.13	0.09	0.25
		0.22	0.56	0.22	1

6. Missing Observations

Returning to the example mentioned in Section 1, combining data from two different pension schemes, we now generate artificially two such datasets by randomly deleting

one or the other of the two covariates in the complete dataset.

6.1. Generating Datasets with Missing Observations

To generate a dataset with missing observations, we first randomly split the complete dataset, containing n individuals, into two subsets, denoted by P_1 and P_2 with sizes n_1 and n_2 respectively ($n_1 + n_2 = n$). We suppose that for each individual in P_1 we observe B but not C, and for each individual in P_2 we observe C but not B. We treat P_1 and P_2 as two separate pension schemes, that we would like to model jointly. Therefore, B and C are never jointly observed in the combined dataset, P_1 combined with P_2. We summarize the features of the combined dataset in Table 6.1. In this section $n_1 = 9,375$ and $n_2 = 9,376$.

Because the combined dataset has been generated from the same source (the complete dataset) there is no further hidden source of heterogeneity, hence the same mortality law can be assumed for all members in P_1 and P_2.

Table 6.1: The set of available data in pension schemes P_1 and P_1.

Sample	**i**	**T**	**X**	**D**	**B**	**C**
P_1	1	t_1	x_1	d_1	b_1	$\times$
	...	...	...	...	...	$\times$
	n_1	t_{n_1}	x_{n_1}	d_{n_1}	b_{n_1}	$\times$
P_2	n_1+1	t_{n_1+1}	x_{n_1+1}	d_{n_1+1}	$\times$	c_{n_1+1}
	...	...	...	...	$\times$	...
	n_1+n_2	$t_{n_1+n_2}$	$x_{n_1+n_2}$	$d_{n_1+n_2}$	$\times$	$c_{n_1+n_2}$

Note: $\times$ indicates the missing observations.

As shown, both covariates are relevant for the analysis, so we would like to take both into consideration, whether observed or not. As mentioned in Section 4.2, when data are missing the likelihood function does not factorize as in equation (4.2), therefore it has to be maximized simultaneously with respect to τ and ζ.

Since the missing variables are categorical, the integral in equation (4.3) is a finite sum over the sample space of the missing variables. For any individual in P_1, whose geo-demographic profile is not observed, the likelihood contribution is:

$$L_i(\tau,\zeta) \propto \sum_{c\in\{0,1,2\}} S_{x_i}(t_i \mid b_i, c; \tau)\, \mu^{d_i}_{x_i+t_i}(b_i, c; \tau)\, f_{B,C}(b_i, c; \zeta) \tag{6.1}$$

while similarly, for those in P_2 we have:

$$L_i(\tau,\zeta) \propto \sum_{b\in\{\text{Low},\text{High}\}} S_{x_i}(t_i \mid b, c_i; \tau)\, \mu^{d_i}_{x_i+t_i}(b, c_i; \tau)\, f_{B,C}(b, c_i; \zeta) \tag{6.2}$$

Hence, in a simplified fashion, the likelihood function is:

$$L\left(\tau,\zeta \mid \mathbf{t},\mathbf{x},\mathbf{b_{obs}},\mathbf{c_{obs}},\mathbf{d}\right) \propto \prod_{\{i\in P_1\}} L_i \prod_{\{i\in P_2\}} L_i \tag{6.3}$$

The unknown paramaters here are: (a) the parameter τ of the hazard function; (b) and the parameter ζ of the joint probability distribution of benefit level and geo-demographic profile. We aim to estimate them jointly. A simplified approach would be to assume fixed values for ζ, and then estimate τ by maximising only $L^{\tau}(\tau)$ from (4.2). Different assumptions about ζ will lead to different estimates of τ, potentially leading to misspecified models and wrong conclusions as discussed in Section 1. However, if a well informed choice for the joint distribution of B and C can be made, this would be a reasonable approach.

6.2. Identifiability

As mentioned in Section 4.2, when data are missing it might not be possible to uniquely estimate the parameters. For this reason, we present two mathematical conditions which need to be met for the identifiability of the parameter vector $\theta = (\tau, \zeta)$.

In this section we discuss the case where neither B nor C is observed for an individual. The resulting likelihood contribution from that individual is that of a finite mixture distribution with the number of components given by the number of possible outcomes of (B, C).

If either B or C is observed, the analysis follows the same arguments with the resulting finite mixture distribution having the same number of components as the number of possible outcomes of the missing variable.

More precisely, identifiability means that for every (τ, ζ) and (τ', ζ') such that $(\tau, \zeta) \neq (\tau', \zeta')$ we have:

$$f_x\left(t;\tau,\zeta\right) \neq f_x\left(t;\tau',\zeta'\right) \qquad \forall\, t, x \tag{6.4}$$

where f_x is defined in (5.3).

In our application, following equation (4.3) the mixture p.d.f. with six components can be written as follows:

$$\begin{aligned} f_x\left(t;\tau,\zeta\right) = {} & \zeta_1 f_x\left(t \mid b = \textsc{Low}, c = 0;\tau\right) + \zeta_2 f_x\left(t \mid b = \textsc{High}, c = 0;\tau\right) \\ & + \zeta_3 f_x\left(t \mid b = \textsc{Low}, c = 1;\tau\right) + \zeta_4 f_x\left(t \mid b = \textsc{High}, c = 1;\tau\right) \\ & + \zeta_5 f_x\left(t \mid b = \textsc{Low}, c = 2;\tau\right) + \zeta_6 f_x\left(t \mid b = \textsc{High}, c = 2;\tau\right) \end{aligned} \tag{6.5}$$

From McLachlan and Peel [2000], the following two conditions are necessary for the identifiability of our finite mixture model.

1. For every value of T and X, different values of (B, C) should return a different p.d.f., that is:

$$f_x\left(t \mid (b,c) = h;\tau\right) \neq f_{x_i}\left(t_i \mid (b,c) = k;\tau\right) \text{ for } h \neq k. \tag{6.6}$$

For a survival model this means that different values of (B, C) return different hazard functions, since this latter characterizes the p.d.f. of T (see equation (2.3));

2. $0 < \zeta_j < 1$ for $j = 1, \ldots, 6$.

Note that while these are necessary conditions, they are not sufficient, see Titterington et al. [1985].

The first condition is needed because if two hazard functions have the same value for different (B, C), then there exist infinitely many values of ζ which return the same value of $f_x(t; \tau, \zeta)$. The second condition is needed because if any value of (B, C) has probability zero, then its related parameters in τ can take any values without changing the mixture probability distribution.

If B and C affect the hazard function multiplicatively, as assumed in our model, the identifiability conditions follow from those of the finite mixtures of Gamma distributions established in Teicher [1963]. In all other cases, in the absence of analytical results the identifiability of our model can be assessed on a case-by-case basis by checking the eigenvalues of the information matrix evaluated at the estimated parameters, as discussed in Section 4.2.

If the two conditions above are fulfilled, the mixture model in (6.5) is identifiable up to a permutation of the parameters, which is sufficient for our purposes.

As mentioned earlier, the extension to the case where either of B or C are observed is straightforward.

6.3. The Fitting Algorithm

When data are missing, maximising the likelihood function may not be straightforward.

- The likelihood with respect to τ and ζ in (6.3) is the sum of $3^{n_1}2^{n_2}$ complete data likelihood functions of the parameters τ and ζ, since for each individual in P_1 and P_2 we sum three and two terms, respectively. Hence, the likelihood function could be multimodal (see Koller and Friedman [2009]).

- A general computational issue is how to handle the constraint on the parameter vector ζ, representing the joint distribution of B and C (its elements are bounded between zero and one, and add up to one).

For these reasons, any constrained optimization routine might suffer from a lack of stability and robustness.

The Expectation-Maximization (EM) algorithm, developed by Dempster et al. [1977] simplifes the estimation task for such problems. However, we found that reformulating the constrained optimisation problem as an unconstrained problem leads to faster convergence and improved robustness of estimates with respect to different starting values of the numerical optimisation routine, compared to an EM algorithm. We apply a Newton-Raphson iterative scheme to the transformed unconditional optimisation problem.

To transform the constraint optimisation problem into an unconstrained problem we apply the Isometric Log-Ratio transform (ILR, Egozcue et al. [2003]) mainly used in

Compositional Data Analysis, for the re-parametrisation of the vector $\zeta = (\zeta_1, ..., \zeta_k)$. The ILR consists of a distance-preserving mapping[3] of the k-dimensional simplex onto $\mathbb{R}^{k-1}$:

$$\pi = \text{ILR}(\zeta) = \mathbf{\Psi}^T \log(\zeta) \tag{6.7}$$

where π is a $(k-1)$-dimensional vector and $\mathbf{\Psi}$ is a matrix of dimension $(k \times (k-1))$, whose columns $(\psi_{\mathbf{1}}, ..., \psi_{\mathbf{k-1}})$ represent the orthonormal basis for the hyperplane of $\mathbb{R}^k$ orthogonal to the k-unit vector $\mathbf{1}$. This means that the block matrix $\left[\mathbf{\Psi}\ \frac{\mathbf{1}}{\sqrt{k}}\right]$ is orthogonal, that is:

$$\left[\mathbf{\Psi};\ \frac{\mathbf{1}}{\sqrt{k}}\right]^T \left[\mathbf{\Psi};\ \frac{\mathbf{1}}{\sqrt{k}}\right] = \mathbf{\Psi}\mathbf{\Psi}^T + \frac{\mathbf{1}\mathbf{1}^T}{k} = \mathbf{I}. \tag{6.8}$$

Following Proposition 1 of Egozcue et al. [2003], the k-dimensional columns, ψ_i, of $\mathbf{\Psi}$ are given by:

$$\psi_i = \sqrt{\frac{i}{i+1}} \left[\underbrace{\frac{1}{i}, ..., \frac{1}{i}}_{\text{i elements}}, -1, \underbrace{0, ..., 0}_{\text{k-i-1 elements}}\right] \tag{6.9}$$

for $i = 1, ..., k-1$.

To obtain the values of ζ for any given π, the inverse of the ILR transform in (6.7) is applied:

$$\zeta = \text{ILR}^{-1}(\pi) = \left[\frac{\exp(\psi_1 \pi)}{\sum_{i=1}^{k} \exp(\psi_i \pi)}, ..., \frac{\exp(\psi_k \pi)}{\sum_{i=1}^{k} \exp(\psi_i \pi)}\right] \tag{6.10}$$

where$(\psi_1, ..., \psi_k)$ are the rows of $\mathbf{\Psi}$.

We now apply the ILR transform to turn our optimisation problem into an unconstrained problem. Our optimisation algorithm then consists of the following steps:

1. Set starting values for the parameter vectors $\widehat{\tau}^{(\mathbf{0})}$ and $\widehat{\zeta}^{(\mathbf{0})}$;
2. Calculate $\widehat{\pi}^{(\mathbf{0})}$ from $\widehat{\zeta}^{(\mathbf{0})}$ using the ILR transform in (6.7):

$$\widehat{\pi}^{(0)} = \text{ILR}\left(\widehat{\zeta}^{(0)}\right); \tag{6.11}$$

[3] **Definition - Distance-preserving mapping (Isometry):** Let X and Y be metric spaces with metrics $\rho_X(\cdot,\cdot)$ and $\rho_Y(\cdot,\cdot)$ respectively. A map $f : X \to Y$ is distance preserving (or isometric) if $\forall x_1, x_2 \in X$ we have $\rho_Y(f(x_1), f(x_2)) = \rho_X(x_1, x_2)$ (see Kolmogorov and Fomin [1975]).

3. In the log-likelihood function, replace ζ_j with the individual components:

$$\frac{\exp(\psi_j \pi)}{\sum_{i=1}^{k} \exp(\psi_i \pi)}, \qquad j = 1, \ldots, k$$

of the inverse ILR transform, see (6.10);

4. Calculate $\widehat{\tau}$ and $\widehat{\pi}$:

$$(\widehat{\tau}, \widehat{\pi}) = \underset{(\tau,\pi)}{\arg\max} \log L\left(\tau, \mathrm{ILR}^{-1}(\pi)\right) \tag{6.12}$$

using a Newton-Raphson algorithm with starting values $\widehat{\tau}^{(\mathbf{0})}$ and $\widehat{\pi}^{(\mathbf{0})}$;

5. Use the inverse ILR transform in (6.10) to calculate $\widehat{\zeta}$ from $\widehat{\pi}$:

$$\widehat{\zeta} = \mathrm{ILR}^{-1}(\widehat{\pi}). \tag{6.13}$$

This is the maximum likelihood estimator for ζ, due to the invariance property of the MLE (see Casella and Berger [2002]).

It is worth emphasizing that possible multimodality means that the solutions of the likelihood equations can be affected by the choice of the starting values. In our numerical study we try several starting values, and choose the MLE which returns the largest value for the log-likelihood function.

6.4. Results of the empirical analysis

To investigate the properties of the estimator $\left(\widehat{\tau}, \widehat{\zeta}\right)$, as mentioned in Section 6.1 we randomly split our original dataset into two parts and remove observations for covariate B from one set and covariate C from the other. In this way, we artificially create the datasets P_1 and P_2 that we then combine and use for the estimation of $\theta = (\tau, \zeta)$. We denote the estimator based on P_1 and P_2 by $\widehat{\theta}_1$.

We first analyse just one scenario of a random split into P_1 and P_2. The results for that scenario are compared with those obtained from the complete dataset and shown in the first two lines of Table 6.2.

Table 6.2: Parameter estimates from the complete dataset ($\widehat{\theta}$), parameter estimates when data are missing for one random split of the complete dataset ($\widehat{\theta}_1$), mean parameter estimate over 1,000 random splits of the complete dataset ($\overline{\widehat{\theta}}$), and standard deviation of the parameters ($\times$ 1,000) over 1,000 random splits of the complete dataset (σ_{PAR}). All figures are rounded to two decimal places.

	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\gamma}$	$\widehat{\delta}_{\mathbf{1}}$	$\widehat{\delta}_{\mathbf{2}}$	$\widehat{\zeta}_{\mathbf{1}}$	$\widehat{\zeta}_{\mathbf{2}}$	$\widehat{\zeta}_{\mathbf{3}}$	$\widehat{\zeta}_{\mathbf{4}}$	$\widehat{\zeta}_{\mathbf{5}}$
$\widehat{\theta}$	−11.41	0.11	−0.21	−0.22	−0.43	0.19	0.03	0.43	0.13	0.13
$\widehat{\theta}_1$	−11.40	0.11	−0.12	−0.27	−0.35	0.22	0.00	0.53	0.03	0.00
$\overline{\widehat{\theta}}$	−11.42	0.11	−0.13	−0.23	−0.33	0.22	0.00	0.53	0.03	0.01
$\sigma_{\mathbf{PAR}} \times 10^3$	16.62	0.19	99.67	30.56	113.29	3.00	0.23	13.80	13.76	13.17

We find in Table 6.2 that when data are missing, the estimators for γ, δ_1, δ_2 and ζ, which depend on B and C, and which should be jointly considered, lead to estimated values that are very different from those obtained when data are fully observed. In particular, $\widehat{\zeta}$ lies on the boundary of the parameter space. If the true value of ζ was on the boundary of the parameter space, then one regularity condition of the maximum likelihood estimator would be violated[4]. Nevertheless, the marginal probability distributions of B and C estimated from data with missing observations are always very close to those estimated from complete data, e.g. $P[B = \text{low}] = \zeta_1 + \zeta_3 + \zeta_5 = 0.75$, which is the same in the first line in Table 6.2 (complete data) and the second line (missing data).

The estimated values of α and β in $\widehat{\theta}_1$ which are estimated using all units in the combined sample P_1 and P_2 are very close to the estimates in $\widehat{\theta}$ obtained from the complete dataset.

We repeat the random splitting of the complete dataset into sets P_1 and P_2, and the estimation exercise 1,000 times. In this way we obtain a sample of 1,000 parameter estimates $\widehat{\theta}_k$, $k = 1, \ldots, 1,000$. The last two rows in Table 6.2 show the average of $\widehat{\theta}_k$, $\overline{\widehat{\theta}} = \frac{1}{1{,}000} \sum_k \widehat{\theta}_k$ and the standard deviation of the parameters over the splits, that is $\sigma_{\text{PAR}} = \sqrt{\frac{1}{999} \sum_k \left(\widehat{\theta}_k - \overline{\widehat{\theta}}\right)^2}$. The numerical results confirm our observations based on the single estimate $\widehat{\theta}_1$.

To investigate the reasons for the poor estimates obtained from P_1 and P_2 combined we consider the identifiability of the parameters. As the missing observations in our study are simulated and the complete dataset is available, we are able to consistently estimate the parameters, which allows us to check whether the identifiability conditions discussed in Section 6.2 are met. Given the hazard function we need to consider the value of $\phi = \exp\left[\gamma \mathbb{1}_{[b=\text{High}]} + \delta_1 \mathbb{1}_{c=c_1} + \delta_2 \mathbb{1}_{c=c_2}\right]$, which defines the hazard function for different values of (B, C) given the baseline and age X. Table 6.3 shows the value of $\widehat{\phi}$ for different values of (B, C) based on the complete dataset.

[4]The true values of the parameter must lie in the interior of the parameter space.

Table 6.3: Estimated values of ϕ for the complete dataset.

$\widehat{\phi}$		**Geo-demographic level (C)**		
		0	**1**	**2**
Benefit level (B)	**Low**	1	0.80	0.65
	High	0.81	0.65	0.53

In Table 6.3 we find that an individual with geo-demographic profile 0 and a high benefit level has a very similar mortality profile to that of an individual with low benefit level and geo-demographic profile 1 (given they are the same age). A similar argument can be made for individuals with high benefits and geo-demographic profile 1 compared to individuals with low benefits and geo-demographic profile 2. The identifiability conditions discussed in Section 6.2 are therefore not fulfilled.

The identifiability of parameters in our model given the available data can be further analysed by looking at the eigenvalues of the Hessian matrix resulting from the estimation process, as discussed in Section 4.2. More precisely, we can investigate whether for the simulated datasets the model behaves as parameter redundant, and hence parameters are not identifiable. While we do not report eigenvalues for all 1,000 simulated datasets in detail, we found that in all of them the two largest eigenvalues of the Hessian matrix are close to zero, indicating that the log-likelihood function has a surface with a flat ridge. This means that the estimated information matrix does not have full rank, and that the model is parameter redundant given the data.

Furthermore, there might be an issue with model risk. We are fitting a Gompertz-type model to our data, which may not reflect the true hazard function for this mortality experience. It is also possible that mortality differentials due to different socio-economic characteristics do not affect the log-hazard proportionally.

In conclusion, given the true value of the parameters, the missing data pattern may lead to a situation which Cole et al. [2010] define as "near parameter redundancy"; that is, the fitting of a full rank statistical model, which behaves as parameter redundant in a specific situation. The same authors argue that this is the consequence of an"inevitably-imprecise numerical method, because the model is very similar to a parameter redundant model for a particular dataset".

7. Availability of Further Information

The identifiability problem arising when two covariates are never jointly observed blocks any further inference on the hazard function. We now investigate the effect of the availability of a further dataset, P_3, of smaller sample size n_3, representing a pension scheme where no covariate information is missing.

In order to show the robustness of our results, we further consider the availability of a dataset, P_4 of larger sample size n_4, representing a pension scheme where both B and C are missing. Table 7.1 illustrates the missing data pattern for P_3 and P_4.

Table 7.1: The set of available data in pension schemes P_3 and P_4.

Sample	i	T	X	D	B	C
	1	t_1	x_1	d_1	b_1	c_1
P_3	...	...	...	...	...	...
	n_3	t_{n_3}	x_{n_3}	d_{n_3}	b_{n_3}	c_{n_3}
	n_3+1	t_{n_3+1}	x_{n_3+1}	d_{n_3+1}	$\times$	$\times$
P_4	...	...	...	...	$\times$	$\times$
	n_3+n_4	$t_{n_3+n_4}$	$x_{n_3+n_4}$	$d_{n_3+n_4}$	$\times$	$\times$

Note: $\times$ indicates the missing observations

We now have four pension schemes, P_1, P_2, P_3 and P_4, obtained by randomly splitting the complete dataset and deleting covariates as appropriate. In the new combined dataset with missing observations, 5% have completely observed data (P_3), 20% have observation of B only (P_1), 20% have observation of C only (P_2), and we fail to observe both B and C for the remaining 55% (P_4) of the units. Hence, $n_1 = 937$, $n_2 = n_3 = 3,748$ and $n_4 = 10,308$. The random splitting of the complete dataset into four pension schemes is again repeated 1,000 times.

The likelihood contribution for each life in P_3 (no missing data) is:

$$L_i(\tau,\zeta) \propto S_{x_i}(t_i \mid b_i, c_i; \tau)\, \mu^{d_i}_{x_i+t_i}(b_i, c_i; \tau)\, f_{B,C}(b_i, c_i; \zeta) \tag{7.1}$$

while for individuals in P_4 (B and C both missing) the likelihood contribution is:

$$L_i(\tau,\zeta) \propto \sum_{b\in\{\text{Low,High}\}} \sum_{c\in\{0,1,2\}} S_{x_i}(t_i \mid b, c; \tau)\, \mu^{d_i}_{x_i+t_i}(b, c; \tau)\, f_{B,C}(b, c; \zeta)\,. \tag{7.2}$$

We compare estimated parameter values and standard errors obtained from different datasets. We consider estimates based on all four sets combined (denoted by $P_1 - P_4$), estimates based on P_1, P_2 and P_3 combined (denoted by $P_1 - P_3$), i.e. excluding only the (large) dataset where both B and C are not observed, and estimates based only on P_3 (no missing data, but very small sample size). The results are shown in Table 7.2.

We compare these estimates with those obtained from the complete dataset (the first line in Table 7.2). We remark that standard errors are estimated by using the negative of the inverse empirical information matrix[5].

[5]The estimation process described in Section 6.3 involves (τ, π), hence the standard errors for (τ, ζ) need to be estimated using the delta method for vector valued functions.

Table 7.2: Parameter estimates from the complete dataset, $\widehat{\theta}$; parameter estimates when data are missing for one random split of the complete dataset, $\widehat{\theta}_1$, and their estimated standard errors, $\widehat{\sigma}\left(\widehat{\theta}_1\right)$ ($\times$ 1,000); mean parameter estimate over 1,000 random splits of the complete dataset, ($\overline{\widehat{\theta}} = \frac{1}{1{,}000}\sum_k \widehat{\theta}_k$); mean of the parameter standard error estimate, ($\overline{\widehat{\sigma}}\left(\widehat{\theta}\right) = \frac{1}{1{,}000}\sum_k \widehat{\sigma}\left(\widehat{\theta}_k\right) \times$ 1,000); and standard deviation ($\times$ 1,000) of the parameter over 1,000 random splits of the complete dataset, ($\sigma_{\text{PAR}} = \sqrt{\frac{1}{999}\sum_k \left(\widehat{\theta}_k - \overline{\widehat{\theta}}\right)^2}$). All figures are rounded to two decimal places.

Dataset	**Estimation**	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\gamma}$	$\widehat{\delta}_{\mathbf{1}}$	$\widehat{\delta}_{\mathbf{2}}$	$\widehat{\zeta}_{\mathbf{1}}$	$\widehat{\zeta}_{\mathbf{2}}$	$\widehat{\zeta}_{\mathbf{3}}$	$\widehat{\zeta}_{\mathbf{4}}$	$\widehat{\zeta}_{\mathbf{5}}$
	$\widehat{\theta}$	−11.41	0.11	−0.21	−0.22	−0.43	0.19	0.03	0.43	0.13	0.13
	$\widehat{\theta}_1$	−11.43	0.11	−0.20	−0.20	−0.27	0.20	0.02	0.43	0.13	0.12
	$\widehat{\sigma}\left(\widehat{\theta}_1\right) \times 10^3$	135.50	1.71	68.34	61.75	83.86	6.85	4.34	9.41	7.81	7.47
$P_1 - P_4$	$\overline{\widehat{\theta}}$	−11.45	0.11	−0.18	−0.19	−0.36	0.19	0.03	0.43	0.13	0.13
	$\overline{\widehat{\sigma}}\left(\widehat{\theta}\right) \times 10^3$	136.43	1.72	68.33	61.66	82.09	6.94	4.68	9.45	7.96	7.66
	$\sigma_{\mathbf{PAR}} \times 10^3$	24.48	0.31	50.42	47.71	64.19	6.22	4.54	8.45	7.52	7.21
	$\widehat{\theta}_1$	−11.39	0.11	−0.24	−0.23	−0.32	0.20	0.02	0.43	0.13	0.12
	$\widehat{\sigma}\left(\widehat{\theta}_1\right) \times 10^3$	200.91	2.55	74.66	67.34	91.44	6.85	4.36	9.41	7.82	7.48
$P_1 - P_3$	$\overline{\widehat{\theta}}$	−11.43	0.11	−0.20	−0.21	−0.41	0.19	0.03	0.43	0.13	0.13
	$\overline{\widehat{\sigma}}\left(\widehat{\theta}\right) \times 10^3$	200.57	2.54	73.99	66.44	88.42	6.95	4.70	9.45	7.96	7.67
	$\sigma_{\mathbf{PAR}} \times 10^3$	148.13	1.88	59.03	55.80	74.72	6.23	4.56	8.46	7.53	7.22
	$\widehat{\theta}_1$	−11.57	0.11	−0.33	−0.33	−0.44	0.20	0.03	0.40	0.14	0.12
	$\widehat{\sigma}\left(\widehat{\theta}_1\right) \times 10^3$	596.78	7.55	156.58	144.24	188.48	13.14	5.37	16.00	11.26	10.60
P_3	$\overline{\widehat{\theta}}$	−11.47	0.11	−0.21	−0.23	−0.44	0.19	0.03	0.43	0.13	0.13
	$\overline{\widehat{\sigma}}\left(\widehat{\theta}\right)$	603.84	7.65	166.74	151.87	198.40	12.81	5.14	16.15	11.08	11.01
	$\sigma_{\mathbf{PAR}} \times 10^3$	597.79	7.59	153.31	150.84	186.81	12.41	5.02	15.82	10.76	10.63

Figure 7.1 shows the empirical c.d.f. of parameters estimates obtained from the 1,000 random splittings of the complete dataset. The two vertical lines indicate the estimated parameter for the case that data are completely observed, $\widehat{\theta}$, and the average estimate using datasets $P_1 - P_4$ combined, that is $\overline{\widehat{\theta}}$.

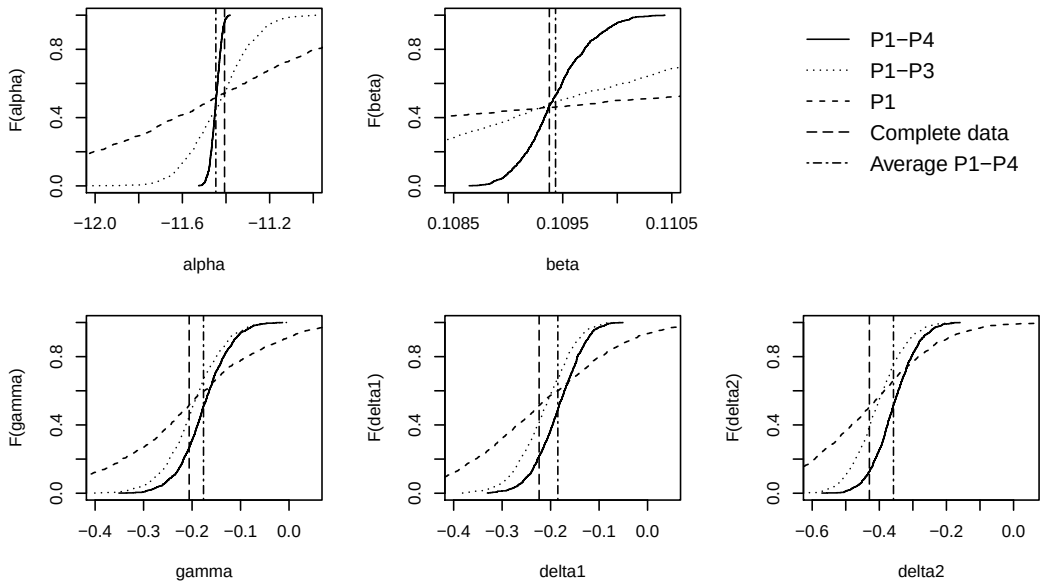

Figure 7.1: Marginal CDFs of the components of $\widehat{\tau}_k$ on the basis of 1,000 random splits of the complete dataset and the dataset combinations.

We find that parameter estimates are in line with estimates obtained based on the complete dataset, as can be seen from the values of $\widehat{\theta}_k$ and $\overline{\widehat{\theta}}$. However, as parts of the complete dataset are excluded from the inference, the estimated standard errors of the parameters $\widehat{\sigma}\left(\widehat{\theta}_k\right)$ increase due to the smaller sample size of the resulting utilized dataset, even if information is incomplete for the excluded individuals.

The increase in the estimated standard errors is most evident for the parameters γ, δ_1 and δ_2 since mortality differentials based on benefit and geo-demographic profile depend on covariates which are sometimes missing. This effect is less remarked for the parameters α and β as age is observed for all individuals, and the baseline is common for all individualss in the four simulated pension scheme datasets. Those differences are confirmed further when comparing the estimated standard errors of $\widehat{\gamma}$, $\widehat{\delta_1}$ and $\widehat{\delta_2}$ when using all four pension schemes ($P_1 - P_4$) with estimates based on pension schemes $P_1 - P_3$ only.

When missing data do not lead to parameter redundancy, the maximum likelihood estimator has its usual asymptotic properties, see Takai and Kano [2013]. Therefore, statistical tests such as the Likelihood-ratio test for restrictions on the parameters, and information criteria such as the AIC and BIC, can still be applied for model selection. However, when data are missing the sample variances might be larger, and statistical tests might have less power.

As in the previous section, we check whether the statistical model behaves as parameter redundant by examining the eigenvalues of the Hessian matrix. We observe that in

all 1,000 missing data scenarios the eigenvalues are all negative and far from zero. In fact, we find that the eigenvalue closest to 0 across all simulations is equal to -9.93.

8. Financial Impact of the Use of Missing Data Statistical Techniques

8.1. Annuity Factor and Mortality Differentials

Let Y_x be the random variable representing the present value of the cash flows a pension fund member aged x receives during his remaining lifetime. If cash flows are assumed to be £1 per year paid continuously, the expected value of Y_x, called the annuity factor, is denoted by $\overline{a}_x$ and calculated as follows (see Dickson et al. [2013]):

$$\overline{a}_x = \mathbb{E}\left(Y_x\right) = \int_0^{120-x} \exp\left(-rt\right) S_x\left(t\right) \mathrm{d}t \tag{8.1}$$

where r is the interest rate (with continuous compounding), 120 is the assumed maximum age, and $S_x\left(t\right)$ is the survival function that depends on the hazard function, as discussed in Section 2. The variance of Y_x is calculated as follows (Dickson et al. [2013]):

$$\mathbb{V}\left(Y_x\right) = \frac{{}^2\overline{A}_x - \overline{A}_x^2}{\delta^2} \tag{8.2}$$

where:

$$\overline{A}_x = \int_0^{120-x} \exp\left(-rt\right) f_x\left(t\right) \mathrm{d}t; \qquad {}^2\overline{A}_x = \int_0^{120-x} \exp\left(-2rt\right) f_x\left(t\right) \mathrm{d}t. \tag{8.3}$$

As we assume a parametric form for the hazard function indexed by the unknown parameter vector τ, any empirical estimate of the variance $\mathbb{V}\left(Y_x\right)$ is affected by the sampling distribution of $\widehat{\tau}$.

The variance of Y_x induced by the sampling distribution of $\widehat{\tau}$ can be decomposed applying the Conditional Variance Identity (Casella and Berger [2002]):

$$\mathbb{V}\left(Y_x\right) = \mathbb{E}\left[\mathbb{V}\left(Y_x \mid \widehat{\tau}\right)\right] + \mathbb{V}\left[\mathbb{E}\left(Y_x \mid \widehat{\tau}\right)\right] \tag{8.4}$$

The second term on the RHS of equation (8.4), $\mathbb{V}\left[\mathbb{E}\left(Y_x \mid \widehat{\tau}\right)\right]$, is helpful to understand the variability of the annuity factor induced by the sampling distribution of $\widehat{\tau}$.

Due to missing data, $\widehat{\tau}$ depends on $\widehat{\zeta}$, hence we need to consider the sampling distribution of $\widehat{\theta} = \left(\widehat{\tau}, \widehat{\zeta}\right)$. Indeed, as hinted in Section 6.4, $\widehat{\theta}$ has asymptotically a multivariate Normal distribution with mean θ and variance-covariance matrix $\mathbf{\Sigma}$. Once again, we estimate the latter using the negative of the inverse Fisher information matrix.

We compare the values of $\overline{a}_x = \mathbb{E}\left(Y_x \mid \widehat{\tau}\right)$ calculated using parameters: (a) estimated by observing all four schemes ($P_1 - P_4$); and (b) estimated by observing schemes $P_1 - P_3$ only. We also compare both results with model M_0 from Section 5.4, which is a simple Gompertz model with no covariates B and C.

The numerical analysis is carried out for two interest rates, $r = 1\%$ and $r = 3\%$, and two attained ages, $x = 65$ and $x = 71$, where the former is the typical retirement age, while the latter is approximately the average age among individuals alive in the complete dataset at the end of the observation period. Table 8.1 shows the values of the annuity factor, the percentage change compared to annuity factors based on model M_0, and the value of $\mathbb{V}\left[\mathbb{E}\left(Y_x \mid \widehat{\tau}\right)\right]$, based on 10,000 sampled values for $\widehat{\theta}$, denoted by θ', which are computed as follows:

$$\theta' = \widehat{\theta} + \mathbf{Au} \tag{8.5}$$

where $\mathbf{A}$ is the lower triangular matrix obtainable from the Cholesky decomposition of $\widehat{\mathbf{\Sigma}}$ (that is $\widehat{\mathbf{\Sigma}} = \mathbf{AA^T}$), and $\mathbf{u}$ is a vector of i.i.d. standard normal variates of the same dimension as θ.

Table 8.1: Values, percentage change and variance of the annuity factor estimated on the basis of an age-only mortality model (M_0 for the whole sample, see Section 5.4) and compared to the model which adjusts mortality for the effect of socio-economic factors estimated based on the four combined pension schemes $P_1 - P_4$, and based on $P_1 - P_3$.

Annuity factor analysis

Age		**rate=1%**			**rate=3%**		
Datasets	**Segmentation**	$\widehat{\ddot{\mathbf{a}}}_{\mathbf{x}}$	**% change**	$\mathbb{V}\left[\mathbb{E}\left(\mathbf{Y_x} \mid \widehat{\theta}\right)\right]$	$\widehat{\ddot{\mathbf{a}}}_{\mathbf{x}}$	**% change**	$\mathbb{V}\left[\mathbb{E}\left(\mathbf{Y_x} \mid \widehat{\theta}\right)\right]$
	Age 65						
$P_1 - P_4$	No segmentation (M_0)	16.77	–	0.007	13.70	–	0.003
$P_1 - P_4$	Low ben., Geo-d. 0	15.49	−7.67	0.07	12.82	−6.44	0.04
	High ben., Geo-d. 0	16.69	−0.49	0.22	13.65	−0.38	0.10
	Low ben., Geo-d. 1	16.66	−0.71	0.03	13.62	−0.57	0.02
	High ben., Geo-d. 1	17.88	6.59	0.14	14.44	5.43	0.06
	Low ben., Geo-d. 2	17.11	2.01	0.15	13.93	1.69	0.07
	High ben., Geo-d. 2	18.34	9.35	0.15	14.75	7.66	0.07
$P_1 - P_3$	Low ben., Geo-d. 0	15.44	−7.97	0.09	12.78	−6.71	0.05
	High ben., Geo-d. 0	16.85	0.44	0.29	13.75	0.38	0.13
	Low ben., Geo-d. 1	16.82	0.25	0.05	13.73	0.22	0.02
	High ben., Geo-d. 1	18.25	8.82	0.19	14.69	7.22	0.08
	Low ben., Geo-d. 2	17.32	3.28	0.20	14.07	2.72	0.09
	High ben., Geo-d. 2	18.77	11.91	0.20	15.03	9.69	0.08
	Age 71						
$P_1 - P_4$	No segmentation (M_0)	12.98	–	0.007	11.02	–	0.004
$P_1 - P_4$	Low ben., Geo-d. 0	11.79	−9.20	0.06	10.14	−8.04	0.04
	High ben., Geo-d. 0	12.90	−0.68	0.19	10.96	−0.57	0.10
	Low ben., Geo-d. 1	12.86	−0.95	0.03	10.93	−0.80	0.02
	High ben., Geo-d. 1	14.00	7.84	0.12	11.76	6.73	0.06
	Low ben., Geo-d. 2	13.29	2.31	0.13	11.25	2.02	0.07
	High ben., Geo-d. 2	14.44	11.20	0.13	12.08	9.56	0.07
$P_1 - P_3$	Low ben., Geo-d. 0	11.75	−9.51	0.08	10.11	−8.32	0.05
	High ben., Geo-d. 0	13.05	0.48	0.25	11.07	0.42	0.13
	Low ben., Geo-d. 1	13.02	0.24	0.04	11.05	0.22	0.02
	High ben., Geo-d. 1	14.36	10.60	0.17	12.02	9.05	0.08
	Low ben., Geo-d. 2	13.49	3.89	0.17	11.39	3.36	0.09
	High ben., Geo-d. 2	14.85	14.36	0.18	12.37	12.20	0.09

The results in Table 8.1 show that for the two attained ages and the two interest rates, there are significant differences between the annuity factors obtained from our model based on socio-economic covariates and the values obtained from Model M_0. Those differences are larger for the lower interest rate, reflecting the fact that mortality is a more important risk factor for annuities when interest rates are low, see Karabey et al. [2014] for a more detailed discussion. As noticed by Richards [2008] the pricing margin for an annuity is around 5%, hence the mortality differentials are extremely material from a financial and business perspective.

We also find in Table 8.1 that relative differences are higher for the older population ($x = 71$). This might actually be a disadvantage of the model applied here as our model M_3 in Section 5.4 does not include an interaction term between age and either of the covariates, B or C. Therefore, one of our implicit modelling assumptions is that mortality differentials between socio-economic groups do not change with age. However, as we are focusing on the effect of missing observations and combining datasets, analysing the effect of a possible interaction term is beyond the scope of the this paper.

In addition, the exclusion of the pension scheme, P_4, with no observed socio-economic covariates leads to higher standard errors of parameter estimates, and, in turn, to an increased variance of annuity factors.

8.2. Mis-estimation Risk Capital Requirement

We now extend our analysis to a portfolio of annuities with different initial ages x and investigate the impact of parameter uncertainty on the capital requirement associated with such a portfolio. The question of interest is whether we can reduce the capital requirement for a given pension scheme by combining its data with data from another pension schemes for the purpose of estimating the parameters of the hazard rate. As an example, we take scheme P_1 (B is observed, but not C) and calculate its capital requirement, and then repeat the process using the combined observations from schemes $P_1 - P_4$ to estimate the hazard rate. Of course, this affects only that part of the total capital requirement related to parameter mis-estimation risk.

Following Richards [2016] for the calculation of the SCR, we perform repeated valuations of the whole annuity portfolio on the basis of a large set of alternative parameter values. The portfolio value as a function of the parameter θ is denoted by $P(\theta)$.

For our portfolio the capital requirement resulting from the risk of mis-estimated parameters is then calculated as follows:

$$\left(\frac{99.5^{th}\text{percentile of } P(\theta)}{\text{mean of } P(\theta)} - 1 \right) \times 100. \tag{8.6}$$

To obtain estimates of the quantile and the mean of $P(\theta)$, we apply the approach in (8.5) to generate 10,000 simulated values for θ. We denote the simulated values again by θ' and obtain that the k-th simulated value of the portfolio, $P(\theta'_k)$ is given by:

$$P\left(\theta'_k\right) = \sum_{i=1}^{n} w_i \overline{a}_{x_i}\left(\theta'_k\right) \tag{8.7}$$

where w_i represents the pension amount of the i-th individual, and where we have emphasised the dependence of the annuity factor on the parameter values.

Based on (8.6) we calculate the mis-estimation risk capital requirement for pension scheme P_1 in three ways.

- The hazard function parameter τ is estimated using only the data available in P_1. In order to keep things simple, and avoid the identifiability issues due to missing C, the hazard function is that of model M_1 (see Section 5.3), whose parameters can be estimated by ordinary maximum likelihood.
- All parameters of the hazard function in (5.1) are estimated using the combined schemes $P_1 - P_3$, using the approach to missing data described in Section 7.
- The parameters in (5.1) are then estimated again using the data from all four schemes $P_1 - P_4$.

The capital requirements for mis-estimation risk based on these two approaches, and the same two interest rates as in Section 8.1, are shown in Table 8.2.

Table 8.2: Mis-estimation capital requirement for pension scheme P_1 on the basis of two interest rate assumptions and on the schemes used to estimate the hazard rates.

	Samples					
	$\mathbf{P_1 - P_4}$		$\mathbf{P_1 - P_3}$		$\mathbf{P_1}$	
Int. rate	1%	3%	1%	3%	1%	3%
Cap. Req.	2.12%	1.68%	2.69%	2.11%	3.96%	3.09%

We repeat the same experiment for pension scheme P_1, which is the other scheme where benefit is observable, whose hazard function specification is always that of equation (5.1). Results are in Table 8.3.

Table 8.3: Mis-estimation capital requirement for pension scheme P_3 on the basis of two interest rate assumptions and on the schemes used to estimate the hazard rates.

	Samples					
	$\mathbf{P_1 - P_4}$		$\mathbf{P_1 - P_3}$		$\mathbf{P_3}$	
Int. rate	1%	3%	1%	3%	1%	3%
Cap. Req.	2.37%	1.88%	2.89%	2.29%	7.27%	5.72%

We see that including the other datasets for the estimation the hazard rates can reduce the capital requirements. This is a direct consequence of the results obtained for the annuity factors: the schemes P_1 and P_3 have smaller sample size with respect to the

combined sample. Therefore, parameter uncertainty is high. Including observations for which some or all covariates are missing increases the sample size and reduces the parameter uncertainty. As we have seen in Section 7, this is particularly true for the parameters α and β which are estimated with a greater degree of certainty even when data are missing.

These results show that including samples with missing observations can reduce the sampling variability of estimates and, in turn, reduce capital requirements, and one way of obtaining such extra observations is combining experience data from different pension schemes assuming that the members of those schemes share the same underlying mortality law.

9. Conclusions

In this paper we addressed the problem of combining the data from different mortality experiences, assumed to have the same probabilistic law for the future lifetime, but which have different covariate information. The aim is to keep the covariate information in estimating a parametric survival model, since this generally improves the estimation in each individual experience. This issue has been tackled by treating the covariates that are unobserved as missing data. The same techniques could be used if covariates were sometimes missing even within a mortality experience. The specific context was provided by empirical data from a U.K. pension scheme.

In particular, we discussed conditions for identifiability of the maximum likelihood estimator (Section 6.2), checking for parameter redundant when data are missing (Section 4.2), and we provided a quick and practical algorithm to address the inferential problem (Section 6.3). We found that:

- when data are missing, the statistical model is not always identifiable using maximum likelihood (Section 6.4); and
- obtaining complete data for a relatively small subset of members will allow us to combine data from two or more experiences and avoid identifiability issues (Section 7).

The latter result is of practical relevance, since obtaining complete information for all the members of different pension funds may be difficult or expensive. This may be compared to a well-known problem faced by the CMI. Because it collected data relating to policies rather than lives, the data contained duplicates — one person having several policies. It could not, in general, de-duplicate this data, but from time to time carried out a special investigation to estimate the distribution of the number of duplicates policies, and with this ancillary information was able to improve the quality of its key outputs, namely life tables for the UK industry.

The techniques in this paper may be useful for an actuary when calculating financial quantities of interest based on annuity factors. Our results showed how the impact can be significant from a business perspective (Section 8). These techniques may allow different datasets with equal or similar mortality experience to be combined, increasing

sample size and reducing parameter risk, therefore, reducing capital requirements. Socio-economic covariates such as benefit level and geo-demographic profile are more relevant when interest rates are low.

However, we have not discussed whether it is still possible to make statistical inferences without discarding any relevant data in a situation of parameter redundancy. This can be better addressed within the Bayesian inferential framework, because when the Fisher information matrix is not strictly positive definite, the maximum likelihood estimator does not have an asymptotic normal distribution (Watanabe [2010]). The use of Bayesian statistical techniques for models of this type when dealing with missing data has been left for future work.

10. Acknowledgements

The authors are grateful to Dr. Stephen J. Richards for very helpful comments. Francesco Ungolo, Marcus C. Christiansen, Torsten Kleinow and Angus S. Macdonald acknowledge financial support from the Actuarial Research Centre of the Institute and Faculty of Actuaries through the research programme on "Modelling, Measurement and Management of Longevity and Morbidity Risk".

A. The likelihood of the observable data

Define:

- $\mathbf{Y} = \{y_{i,j}\}$ is a $(n \times p)$ rectangular dataset, where the i-th row $\mathbf{y_i} = (y_{i,1}, ..., y_{i,p})$ represents the realization of the p-dimensional random vector $\mathbf{Y}$ with density $f_{\mathbf{Y}}(\ \cdot\ ; \theta_0)$. In this work $\mathbf{Y} = (\mathbf{T}, \mathbf{X}, \mathbf{Z})$, although we hereby keep things in a general fashion;
- $\mathbf{M} = \{m_{i,j}\}$ is the $(n \times p)$ missing indicator matrix, where the i-th row $\mathbf{m_i} = (m_{i,1}, ..., m_{i,p})$ represents the realization of the p-dimensional random vector $\mathbf{M}$ indicating the missing data probabilistic mechanism, with density $f_{\mathbf{M}|\mathbf{Y}}(\ \cdot\ |\ \cdot\ )$. In particular, $m_{i,j} = 1$ if $y_{i,j}$ is missing and 0 otherwise.

Suppose for each sampled unit only some elements of $\mathbf{y_i} = (y_{i,1}, ..., y_{i,p})$ can be observed. Let $\mathbf{y_{i,obs}}$ be the vector of observables, which is of dimension $p - p'$, where $0 \leq p' \leq p$, while the remaining p' variables are missing in the sense that there exists an underlying value useful for analysis, but that value is unknown. The subvector of missing observations for the i-th individual can be denoted as $\mathbf{y_{i,mis}}$, which is then of dimension p', and depends on the i-th sampled unit.

In case of completely observed and independently distributed data, the i-th individual likelihood contribution is given by:

$$L_i(\theta) = f_{\mathbf{Y}}(\mathbf{y_i}; \theta) \tag{A.1}$$

while if missing data issue can occur throughout an experiment, then the following full density which accounts for the missing data mechanism is to be taken into account:

$$f_{\mathbf{Y},\mathbf{M}}\left(\mathbf{y_i},\mathbf{m_i};\theta\right)=f_{\mathbf{Y}}\left(\mathbf{y_i};\theta\right)f_{\mathbf{M}|\mathbf{Y}}\left(\mathbf{m_i}\mid\mathbf{y_i}\right) \tag{A.2}$$

$\mathbf{M}$ can take at most 2^p values, which can be enumerated as $\mathbf{m^{(0)}},...,\mathbf{m^{(2^p-1)}}$, where $\mathbf{m^{(0)}}=(0,0,...,0)$ represents the case where all elements of $\mathbf{y_i}$ are observed for the i-th individual, while $\mathbf{m^{(2^p-1)}}=(1,1,...,1)$ where all elements of $\mathbf{y_i}$ are missing.

To each $k=0,...,2^p-1$, describing the missing data pattern, there correspond a different set of indices indicating the elements of $\mathbf{y_i}$ which have been observed[6].

At this purpose let us define $\mathbf{y^{(k)}}$ to be the subvector of $\mathbf{y}$, of dimension given by the number of zeroes in $\mathbf{m^{(k)}}$, whose elements are those of index k, while $\mathbf{y^{(-k)}}$ represents the subvector of $\mathbf{y}$ of dimension given by the number of ones in $\mathbf{m^{(k)}}$, and which includes those elements not included in $\mathbf{y^{(k)}}$.

On the base of these definitions, the joint density of $\left(\mathbf{Y^{(k)}},\mathbf{M^{(k)}}\right)$ for the i-th individual is written as follows:

$$f_{\mathbf{Y^{(k)}},\mathbf{M^{(k)}}}\left(\mathbf{y_i^{(k)}},\mathbf{m^{(k)}};\theta\right)=\int_{\mathbb{Y}^{(-k)}}f_{\mathbf{Y}}\left(\mathbf{y_i};\theta\right)f_{\mathbf{M^{(k)}}|\mathbf{Y}}\left(\mathbf{m^{(k)}}\mid\mathbf{y_i}\right)d\mathbf{y_i^{(-k)}} \tag{A.3}$$

where the dimension of the integral is given by the number of zeroes in $\mathbf{m^{(k)}}$, and the extremes of integration are given by the sample space of $\mathbf{Y^{(-k)}}$.

When data are missing at random, then $f_{\mathbf{M}|\mathbf{Y}}\left(\mathbf{m_i}\mid\mathbf{y_i}\right)$ in equation (A.2) and $f_{\mathbf{M^{(k)}}|\mathbf{Y}}\left(\mathbf{m^{(k)}}\mid\mathbf{y_i}\right)$ in equation (A.3) can be treated as multiplicative constant.

Hence, we derive the likelihood contribution for the i-th individual: we first define the (2^p-1)-dimensional random vector $\mathbf{M'}$, observed for all individuals, where the k-th element takes value of 1 if for the i-th individual $\mathbf{m_i}=\mathbf{m^{(k)}}$ and 0 otherwise for $k=1,...,2^p-1$. In case of $\mathbf{m_i}=\mathbf{m^{(0)}}$, then all elements of the observed $\mathbf{M'}$ are equal to zero.

Hence, $\mathbf{m_i'}$ has at most one element equal to 1 and the remaining equal to zero.

Finally, the likelihood contribution of the i-th individual, based observed data $(\mathbf{y_{i,obs}},\mathbf{m_i})$ can be written as:

$$\begin{aligned}L_i\left(\theta\right)&=f_{\mathbf{Y_{obs}}}\left(\mathbf{y_{i,obs}},\mathbf{m_i};\theta\right)\\&=f_{\mathbf{Y}}\left(\mathbf{y};\theta\right)^{\left(1-m'_{i,1}-...-m'_{i,(2^p-1)}\right)}\times f_{\mathbf{Y^{(1)}}}\left(\mathbf{y_i^{(1)}};\theta\right)^{m'_{i,1}}\times\dots\\&\times f_{\mathbf{Y^{(2^p-1)}}}\left(\mathbf{y_i^{(2^p-1)}}\theta\right)^{m'_{i,(2^p-1)}}\end{aligned} \tag{A.4}$$

In the main body of this work we simplify the notation, and write the likelihood contribution based on observed data as follows:

$$L_i\left(\theta\right)=f_{\mathbf{Y_{obs}}}\left(\mathbf{y_{i,obs}};\theta\right)=\int f_{\mathbf{Y}}\left(\mathbf{y_i};\theta\right)d\mathbf{y_{i,mis}} \tag{A.5}$$

[6]This means that k represent a one to one mapping from the set of index combinations of the elements of $\mathbf{y}$ to the set $\{0,1,...,2^p-1\}$

where the integral in (A.5) (as well as those implied in (A.4), is over the sample space of the missing random variable, and its dimension is the same as the dimension of the missing random vector.

In other words, on the base of the observed value of $\mathbf{m_i}$ equation (A.5), and then (A.4) involve the conditional distribution due to the observed elements of $\mathbf{y_i}$.

B. Levels for benefit amount and geo-demographic profile

The geo-demographic profile is available on the base of the eighteen MOSAIC codes shown in Table B.1.

First of all, we group individuals with codes 90, 91 and 92 (Guernsey, Jersey, Isle of Man), which will be uniquely indicated as 9X for reasons: few individuals in the sample have these codes, and these latter refer to off-shore crown dependencies.

In this study we discard exposures of those individuals whose geo-demographic profile is missing and of those with code 98 and 99.

Table B.1: Geodemographic composition for the working dataset.

Group	**Records**	**Group**	**Records**	**Group**	**Records**
A	206	G	285	M	2,049
B	1,236	H	129	N	1,164
C	218	I	2,128	O	413
D	1,523	J	3,105	90	3
E	1,025	K	512	91	1
F	3,272	L	1,466	92	6

Given these preliminary considerations, the grouping has been carried out according to the following steps:

1. Fit a model with α and β only as in model M_0 for the whole population;

2. For each sub-population with the same geo-demographic profile (*geo*), after fixing β at the value obtained in Step 1, we fit the following model:

$$\mu_{t_i}(x_i, c_i; \tau) = \exp\left[\alpha_{geo} + \beta\,(x_i + t_i)\right] \tag{B.1}$$

   In this way we obtain 16 different estimates of α[7];

3. Group the standardized values of $\widehat{\alpha}_{geo}$ on the base of a minimum distance criterion. For example[8] we can produce the dendrogram of Figure B.1, which provides

[7] β is fixed because its MLE is strongly correlated with the MLE of α. For example, if in a subsample we have very old individuals, then β will be very low and α very high. Hence in the comparisons of the mortality rates due to different geo-demographic profiles, also the choice of x becomes crucial, which is something we prefer to avoid to keep things simple.

[8] Other clustering criteria, such as K-means, might produce a slightly different clustering

a hierarchical graphical inspection of the groupings (obtained by the method of Ward Jr. [1963])

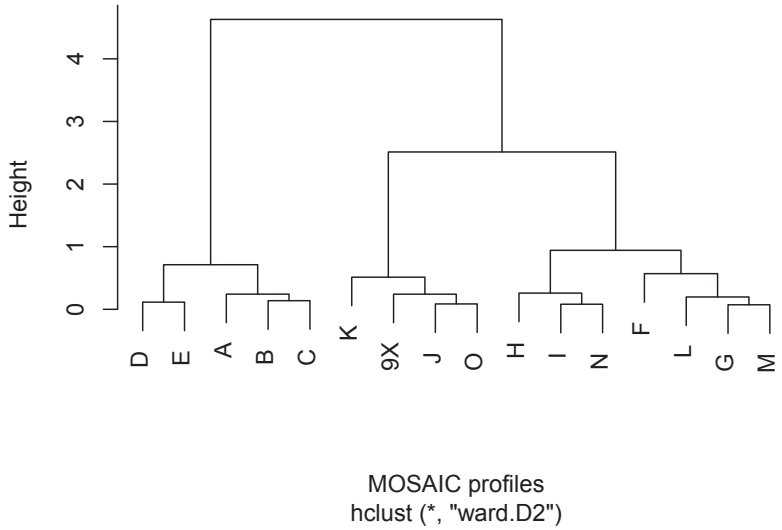

Figure B.1: Cluster dendrogram of geo-demographic profiles.

Table B.2 shows the standardized value of $\widehat{\alpha}_{geo}$:

Table B.2: Standardized values of fitted α_{geo}.

Code	**std**$(\widehat{\alpha}_{\mathbf{geo}})$	**Code**	**std**$(\widehat{\alpha}_{\mathbf{geo}})$	**Code**	**std**$(\widehat{\alpha}_{\mathbf{geo}})$
A	-1.258	G	0.138	M	-0.066
B	-1.399	H	0.288	N	0.553
C	-1.537	I	0.472	O	1.268
D	-0.881	J	1.183	9X	1.016
E	-0.995	K	1.575		
F	-0.420	L	-0.069		

For the choice of the optimal number of clusters plenty of indexes are available, although in any case each of them depends on the adopted clustering technique. For this reason, in order to keep things simple, and focus on the treatment of missing data, we chose for simplicity only three groupings:

- Geo-demographic level 0: J, K, O, 9X;
- Geo-demographic group 1: F, G, H, I, L, M, N;
- Geo-demographic group 2: A, B, C, D, E.

The same methodology has been adopted for the benefit covariate. We divide the whole population into four sub-populations on the base of the benefit amount quartiles, obtaining the dendrogram of Figure B.2.

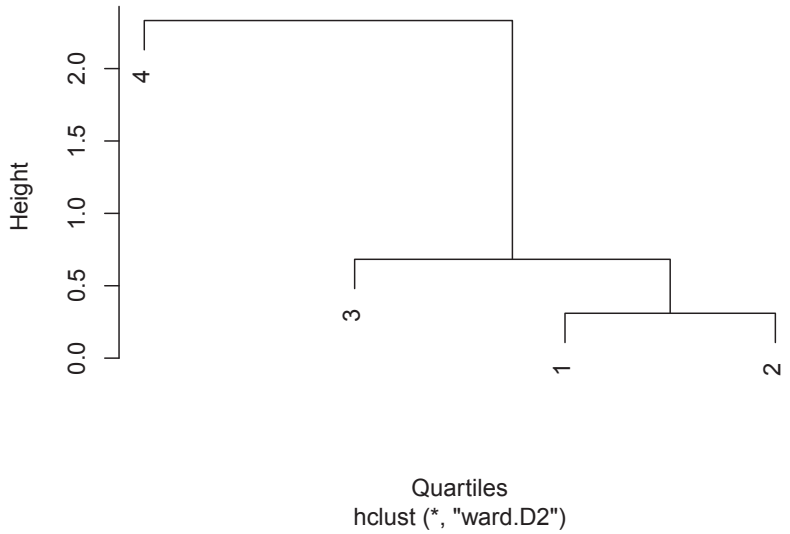

Figure B.2: Cluster dendrogram of benefit quartiles.

Again, we show also the standardized values of $\widehat{\alpha}_q$ for the four quartiles q:

Table B.3: Benefit quartile bands and standardized values of fitted α_q.

Quartile	**Benefit band**	**std**$(\widehat{\alpha}_{\mathbf{q}})$
1^{st}	$0 - 2,376.65$	0.518
2^{nd}	$2,376.66 - 4,811.97$	0.828
3^{rd}	$4,811.98 - 8,553.09$	0.082
4^{th}	$8,553.10 - 116584.44$	-1.428

Hence, given the split suggested by the dendrogram, which separates the individuals in the upper quartiles from all the others, we chose a threshold of £8,500. This implies that 14,003 individuals (74.62% of the pension scheme population) will be treated as low benefit pensioners, and the remaining 4,738 (25.38% of the pension scheme population) as high benefit pensioners.

C. Exposures and Deaths for the available dataset

Table C.1: Exposure and deaths for the dataset split by socio-economic characteristics.

Age last	Population		Low benefit		High benefit		Geo-dem. 0		Geo-dem. 1		Geo-dem. 2	
birthday	Exposures	Deaths	Exposures	Deaths	Exposures	Deaths	Exposures	Deaths	Exposures	Deaths	Exposures	Deaths
60	8294.40	75	5725.47	60	2568.94	15	1792.07	25	4632.98	40	1869.35	10
61	8352.14	76	5811.16	58	2540.98	18	1821.54	26	4643.01	41	1887.59	9
62	8347.98	78	5864.04	66	2483.94	12	1823.69	28	4616.37	37	1907.93	13
63	8170.40	67	5807.13	54	2363.27	13	1803.69	21	4499.64	37	1867.07	9
64	8216.26	76	5950.71	60	2265.55	16	1822.18	31	4523.46	31	1870.62	14
65	9989.07	105	7584.21	87	2404.86	18	2177.68	34	5604.55	61	2206.84	10
66	9687.87	103	7437.91	88	2249.96	15	2120.70	26	5449.58	63	2117.59	14
67	9364.24	136	7234.68	120	2129.57	16	2058.70	38	5281.77	77	2023.78	21
68	9083.86	126	7050.68	105	2033.18	21	2008.59	29	5113.21	80	1962.06	17
69	8730.38	144	6797.29	127	1933.09	17	1915.72	57	4941.95	66	1872.71	21
70	8376.30	153	6536.77	128	1839.52	25	1835.42	54	4729.56	81	1811.31	18
71	7979.19	178	6245.21	150	1733.99	28	1751.01	62	4497.35	94	1730.84	22
72	7580.31	174	5940.66	144	1639.65	30	1659.44	46	4266.52	104	1654.35	24
73	7152.28	201	5617.63	163	1534.66	38	1567.80	53	4010.22	117	1574.27	31
74	6591.60	190	5177.68	153	1413.92	37	1453.11	50	3695.44	103	1443.05	37
75	6074.43	209	4773.64	174	1300.79	35	1334.23	59	3399.52	118	1340.68	32
76	5573.20	193	4367.91	161	1205.29	32	1210.10	50	3118.85	112	1244.25	31
77	5109.13	209	3994.58	170	1114.55	39	1106.78	62	2847.25	111	1155.10	36
78	4610.55	209	3591.53	172	1019.02	37	992.39	58	2559.79	115	1058.37	36
79	4100.03	216	3186.83	182	913.20	34	883.17	50	2277.41	111	939.45	55
80	3623.83	206	2827.82	166	796.01	40	779.34	50	2018.56	116	825.93	40
81	3138.12	211	2435.35	173	702.77	38	678.73	42	1735.23	129	724.17	40
82	2711.60	194	2087.57	165	624.03	29	588.33	53	1496.84	105	626.43	36
83	2334.00	190	1783.64	155	550.37	35	497.03	51	1290.59	103	546.38	36
84	1977.87	167	1497.71	133	480.16	34	412.99	40	1081.60	99	483.28	28
85	1653.55	157	1236.38	119	417.17	38	334.61	41	905.44	77	413.49	39
86	1357.01	169	1006.05	132	350.96	37	273.10	34	737.66	97	346.25	38
87	1104.61	122	816.39	94	288.22	28	222.94	23	590.02	70	291.65	29
88	874.10	130	637.89	91	236.21	39	164.37	33	468.90	65	240.83	32
89	665.76	114	481.29	86	184.47	28	121.05	20	366.88	61	177.83	33
90	494.15	85	359.77	59	134.38	26	86.43	21	274.39	43	133.32	21
91	381.22	67	275.96	53	105.26	14	63.04	17	206.46	35	111.72	15
92	292.19	59	208.99	38	83.20	21	40.39	8	160.54	38	91.26	13
93	206.40	54	149.64	42	56.76	12	27.80	10	109.35	28	69.25	16
94	148.98	33	108.22	25	40.77	8	20.01	4	78.10	17	50.88	12
95	95.69	29	71.07	18	24.62	11	15.04	3	46.27	14	34.37	12
96	57.66	22	42.15	17	15.51	5	10.99	4	28.46	9	18.21	9
97	37.10	9	25.88	7	11.22	2	5.64	2	18.45	7	13.00	0
98	24.80	6	17.54	2	7.26	4	4.00	0	9.34	5	11.46	1
99	16.93	7	12.27	6	4.66	1	3.35	0	6.06	2	7.52	5
100	9.53	4	7.26	4	2.28	0	2.21	1	4.18	1	3.15	2
101	6.03	2	4.03	2	2.00	0	1.69	1	3.00	0	1.34	1
102	5.00	0	3.00	0	2.00	0	1.00	0	3.00	0	1.00	0
103	1.61	1	1.04	0	0.57	1	0.36	0	1.01	0	0.23	1
Total	172601.39	4956	130792.58	4009	41808.80	947	37492.47	1317	96348.78	2720	38760.14	919

References

G. Casella and R. Berger. *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002.

E. A. Catchpole and B. J. T. Morgan. Detecting parameter redundancy. *Biometrika*, 84 (1):187–196, 1997.

C. CMI. Working paper 26: Extensions to younger ages of the "00" series pensioner tables of mortality. *Continuous Mortality Investigation, London*, 2007.

D. Cole, B. Morgan, and D. Titterington. Determining the parametric structure of models. *Mathematical Biosciences*, 228(1):16–30, 2010.

D. Cox and H. Miller. *The Theory of Stochastic Processes*. Methuen's monographs on applied probability and statistics. Taylor & Francis, 1977. ISBN 9780412151705.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

D. Dickson, M. Hardy, and H. Waters. *Actuarial Mathematics for Life Contingent Risks*. International Series on Actuarial Science. Cambridge University Press, 2013. ISBN 9781107044074.

J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35 (3):279–300, Apr 2003.

B. Gompertz. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society of London*, 115:513–583, 1825.

A. H. Herring and J. G. Ibrahim. Likelihood-based methods for missing covariates in the cox proportional hazards model. *Journal of the American Statistical Association*, 96(453):292–302, 2001.

U. Karabey, T. Kleinow, and A. J. Cairns. Factor risk quantification in annuity models. *Insurance: Mathematics and Economics*, 58:34 – 45, 2014.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

A. Kolmogorov and S. Fomin. *Introductory Real Analysis*. Dover Books on Mathematics. Dover Publications, 1975. ISBN 9780486612263.

F. M. Lord. Estimation of parameters from incomplete data. *Journal of the American Statistical Association*, 50(271):870–876, 1955. ISSN 01621459.

A. S. Macdonald. An actuarial survey of statistical models for decrement and transition data, i: multiple state, poisson and binomial models. *British Actuarial Journal*, 2: 129–155, 1996.

A. S. Macdonald, S. J. Richards, and I. D. Currie. *Modelling Mortality with Actuarial Applications*. International Series on Actuarial Science. Cambridge University Press, 2018.

A. M. Madrigal, F. E. Matthews, D. D. Patel, A. T. Gaches, and S. D. Baxter. What longevity predictors should be allowed for when valuing pension scheme liabilities? *British Actuarial Journal*, 16(1):1–38, 2011.

G. J. McLachlan and D. Peel. *Finite mixture models*. Wiley Series in Probability and Statistics, New York, 2000.

S. J. Richards. Applying survival models to pensioner mortality data (with discussion). *British Actuarial Journal*, 14(2):257–326, 2008.

S. J. Richards. Mis-estimation risk: measurement and impact. *British Actuarial Journal*, 21(3):429–457, 2016.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

M. D. Schluchter and K. L. Jackson. Log-linear analysis of censored survival data with partially observed covariates. *Journal of the American Statistical Association*, 84(405): 42–52, 1989. ISSN 01621459.

G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, March 1978.

K. Takai and Y. Kano. Asymptotic inference with incomplete data. *Communications in Statistics - Theory and Methods*, 42(17):3174–3190, 2013.

H. Teicher. Identifiability of finite mixtures. *Ann. Math. Statist.*, 34(4):1265–1269, December 1963.

D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York, 1985.

J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

S. Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.*, 11:3571–3594, Dec. 2010. ISSN 1532-4435.

H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1): 1–25, 1982.

S. S. Wilks. Moments and distributions of estimates of population parameters from fragmentary samples. *Ann. Math. Statist.*, 3(3):163–195, 08 1932. doi: 10.1214/aoms/1177732885.